League tables for literacy survey data based on random effect models

Nick Sofroniou[1], Dominique Hoad[2] and Jochen Einbeck[2]

[1] Educational Research Centre, St Patrick's College, Dublin, Ireland, nick_sofroniou@yahoo.co.uk
[2] Durham University, Department of Mathematical Sciences, Durham, UK

Abstract: Data from the International Adult Literacy Survey are used to illustrate how league tables can be obtained from summary data, consisting of percentages and their standard errors, using random effects models estimated by nonparametric maximum likelihood.

Keywords: Random effect models; Mixture models, Nonparametric maximum likelihood; Effective sample sizes.

1 Introduction

In the years 1994–95 twelve countries participated in the International Adult Literacy Survey (IALS). Literacy is defined as *using printed and written information to function in society, to achieve one's goals and to develop one's knowledge and potential*. The IALS developed a rigorous framework, building on work done in the 1985 Young Adult Literacy Survey (YALS) which consisted of three scales: prose, document and quantitative. It was felt that these three scales were the most significant for measuring literacy and sufficiently practical, with speaking and listening being too costly to measure. We concentrate on the measurement of prose in this paper. The data were reported by giving the percentage of individuals achieving prose level $1, 2, \ldots, 5$, with level 1 being worst. One way of analyzing these data is to dichotomize the data around the lowest cutpoint (i.e., the threshold between level 1 and level 2) to give percentages of adults in each country who could/could not reach a basic level of literacy. This is of particular interest to educationalists and policy makers concerned with social inclusion and its educational and economic implications. For the prose measure, the data can then be summarized in the form of Table 1.

2 Methodology for league table construction

2.1 Effective sample sizes

The IALS used complex sample designs that varied with each country and which involved both stratification by factors such as region or school size, and clustering of pupils within schools. This complicates the issue of the

TABLE 1. Percentage of adults not reaching at least Level 2 for 12 countries (Switzerland was split into two parts according to language and is treated as if it were two separate countries. Canada was treated as English-speaking and Belgium (Flanders) as Dutch). SE denotes the standard error of this percentage and n the sample size.

	Male			Female		
Country	n	% Level 1	SE	n	% Level 1	SE
Sweden	1289	7.31	0.80	1355	7.18	1.03
Netherlands	1358	10.39	1.08	1479	10.49	0.98
Germany	938	14.31	1.89	1124	13.31	1.85
Australia	3767	18.33	0.85	4437	15.69	0.78
Canada	1979	18.76	2.03	2521	14.44	2.04
New Zealand	1821	19.94	1.28	2402	16.52	1.46
Belgium (Flanders)	1066	15.55	1.69	1180	21.61	2.29
Switzerland (French)	682	17.46	1.88	751	19.44	1.70
Switzerland (German)	659	18.30	1.51	733	20.66	1.66
United Kingdom	1730	21.38	1.26	2081	21.60	1.82
Ireland	1050	24.21	2.91	1319	20.93	1.32
United States	1416	23.00	1.65	1577	18.76	1.45
Poland	1431	43.72	0.91	1569	41.74	1.74

denominator to be used in mixed binomial models as the effective sample sizes will tend be considerably less than the actual number of students in each country, due to the intra-cluster correlations.

Cochran (1977) states that under simple random sampling the sample proportion $p = a/n$ is an unbiased estimate of the population proportion $P = A/N$ and that an unbiased estimate of the variance of p obtained from the sample is

$$v(p) = s_p^2 = \frac{N-n}{(n-1)N}pq$$

which simplifies even further assuming large N, and hence a negligible finite population correction, to

$$v(p) = \frac{pq}{n-1}.$$

By rearranging these expressions one can obtain the corresponding sample size n under simple random sampling, e.g.,

$$n = \frac{N(pq + v(p))}{pq + Nv(p)}$$

for the former expression. Thus, it is possible to use the summary information in Table 1, consisting of percentages and their standard errors, to calculate an effective sample size corresponding to the number of independent observations in a theoretical simple random sample. This allows the use of standard mixed binomial modelling software with the effective

sample size as the binomial denominator, reflecting the uncertainty in the percentages of the original table.

2.2 Random effect models

Probabilities are commonly either modelled through a binomial logit or a Poisson log model, with the latter one being less adequate in this example as we have relatively large probabilities and small sample sizes involved. The variability of the upper-level units, here countries i, can be taken into account by adding a random intercept z_i with *unspecified* distribution $g(\cdot)$ to the linear predictor, so that the the binomial random effect model takes the form

$$\log \frac{p_{ij}}{1-p_{ij}} = \beta' x_{ij} + \gamma' s_i + z_i \tag{1}$$

where x_{ij} contains the lower-level covariates (here only `gender`), s_i contains the upper-level covariates (here only the factor for `language` is considered), γ is a non-random parameter and β is a fixed or random parameter. Such variance component models with unspecified random effect distribution can be conveniently fitted using the method of nonparametric maximum likelihood (Aitkin, Hinde & Francis, 2005, p. 440ff), which is implemented in the R package **npmlreg** (Einbeck, Hinde & Darnell, 2007). In short, the density $g(\cdot)$ is approximated by a discrete distribution with K mass points, the locations z_k and masses π_k, $k = 1, \ldots, K$, of which are estimated through the EM algorithm. Thereby, the E-step corresponds to updating of the probabilities $w_{ik} = P(\text{unit } i \text{ comes from mass point } k)$, and the M-step to a weighted generalized linear model fit with weights w_{ik}. From the set of weights after the final EM iteration, one computes posterior intercepts $z_i = \sum_k w_{ik} z_k$ which represent the cluster-level contribution to the response, adjusted by the covariates. As this posterior intercept "sticks" to the cluster for all its lower-level units, it forms a characteristic of the cluster (country). Sofroniou, Einbeck & Hinde (2006) used the posterior intercept for the construction of league tables in the absence of upper-level covariates.

3 Results and conclusions for the literacy survey

We considered several additive logistic random effect models of type (1). To keep the model parsimonious (with only 26 observations available), the models considered exclude a `language.gender` interaction term and random coefficients. Fitting gender as a covariate and no language factor requires 5 masspoints for the random intercept distribution and has a disparity of $-2 \log L = 229.0$ with $df = 16$. Table 2 gives posterior probabilities of the membership of each country to a given component. It suggests that there are two main groups of countries, two countries who performed considerably better (Sweden and the Netherlands), and one low scoring outlier

TABLE 2. Posterior probabilities for the IALS data.

	Posterior intercept	Masspoints				
Intercept		−2.602	−2.156	−1.599	−1.379	−0.330
Proportion		0.077	0.093	0.434	0.319	0.077
Sweden	−2.60	1.00				
Netherlands	−2.16		1.00			
Germany	−1.72		0.21	0.79		
Australia	−1.60			1.00		
Canada	−1.59			0.97	0.03	
New Zealand	−1.58			0.92	0.08	
Belgium (Flanders)	−1.58			0.89	0.11	
Switzerland (French)	−1.54			0.72	0.28	
Switzerland (German)	−1.45			0.34	0.66	
Ireland	−1.38				1.00	
United Kingdom	−1.38				1.00	
United States	−1.38			0.01	0.99	
Poland	−0.33					1.00

Posterior probabilities: $p \geq 0.95$, $0.90 \leq p < 0.95$, $p < 0.90$.

(Poland). Adding `language` dichotomized into English/non-English speaking required 5-masspoints and reduced the disparity to 223.3 with $df = 15$. This was further improved by using all 6 levels of language, with a disparity of 210.3, $df = 15$, and 3 masspoints. However, several categories are based on only a single country and so their performance levels become confounded with language spoken. Therefore, we experimented with adding the fitted upper-level contribution to the posterior intercept, yielding a similar league table to the one presented above, but further research on issues such as representing the uncertainty corresponding to each value is required. These last two models provide some evidence in favour of the suggestion that one contribution to the observed differences in performance may be that of the language of testing.

References

Aitkin, M., Francis, B. and Hinde, J. (2005). *Statistical Modelling in GLIM 4* (2nd edn.). Oxford, UK.

Cochran, W.G. (1977). *Sampling Techniques* (3rd edn.). Wiley, New York.

Einbeck, J., Hinde, J. and Darnell, R. (2007). A new package for fitting random effect models – The npmlreg package. *R News*, **7**, 26–30.

Sofroniou, N., Einbeck, J. and Hinde, J. (2006). Analyzing Irish Suicide Rates with Mixture Models. In *Proceedings of the 21st International Workshop on Statistical Modelling,* Galway, Ireland, 474–481.